Revealing the spatio-phenotypic patterning of cells in healthy and tumor tissues with mLSR-3D and STAPL-3D

Ineveld, R.L. van; Kleinnijenhuis, M.; Alieva, M.; Blank, S. de; Roman, M.B.; Vliet, E.J. van; ... ; Rios, A.C.

nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-00926-3

Check for updates

Revealing the spatio-phenotypic patterning of cells in healthy and tumor tissues with mLSR-3D and STAPL-3D

Ravian L. van Ineveld[1,2,5], Michiel Kleinnijenhuis[1,2,5], Maria Alieva[1,2], Sam de Blank[1,2], Mario Barrera Roman[1,2], Esmée J. van Vliet[1,2], Clara Martínez Mir[1,2], Hannah R. Johnson[1,2], Frank L. Bos[1,2], Raimond Heukers[3], Susana M. Chuva de Sousa Lopes[4], Jarno Drost[1,2], Johanna F. Dekkers[1,2], Ellen J. Wehrens[1,2] and Anne C. Rios[1,2] ✉

Despite advances in three-dimensional (3D) imaging, it remains challenging to profile all the cells within a large 3D tissue, including the morphology and organization of the many cell types present. Here, we introduce eight-color, multispectral, large-scale single-cell resolution 3D (mLSR-3D) imaging and image analysis software for the parallelized, deep learning–based segmentation of large numbers of single cells in tissues, called segmentation analysis by parallelization of 3D datasets (STAPL-3D). Applying the method to pediatric Wilms tumor, we extract molecular, spatial and morphological features of millions of cells and reconstruct the tumor's spatio-phenotypic patterning. In situ population profiling and pseudotime ordering reveals a highly disorganized spatial pattern in Wilms tumor compared to healthy fetal kidney, yet cellular profiles closely resembling human fetal kidney cells could be observed. In addition, we identify previously unreported tumor-specific populations, uniquely characterized by their spatial embedding or morphological attributes. Our results demonstrate the use of combining mLSR-3D and STAPL-3D to generate a comprehensive cellular map of human tumors.

Single-cell resolution volumetric imaging permits the exploration of intact tissues[1–4], retaining spatial and geometrical information that is often lost through tissue dissociation in other single-cell technologies[5]. It thereby has the important advantage of revealing—in a single overview—the relationships between diverse cell types that both normal organ development and cellular function depend on, and how this is shifted under pathological conditions, such as cancer. Despite advances in 3D image processing, including nuclear and membrane segmentation methods[6–11] and large-scale nuclei counting of intact human organs[12], delineating the exact cellular organization of large human tissues at the single-cell level remains a challenge. The high number of cellular subsets and their various morphologies and configurations all complicate accurate single-cell identification and profiling. While challenging, such an approach would be highly informative, as it creates a single-cell readout that retains spatial and morphometric information and can thereby phenotype cells in the context of their native tissue environment. Therefore, to fully exploit the potential of volumetric imaging, we here developed multispectral large-scale single-cell resolution 3D (mLSR-3D) imaging with 'on-the-fly' linear unmixing for single-scan acquisition of eight spectrally resolved fluorophores. Combined with segmentation analysis by parallelization of 3D datasets (STAPL-3D), an automated pipeline for compartment-specific feature extraction, it enables in situ analysis of millions of cells in tissue (Fig. 1a and Supplementary Video 1).

Results

To interrogate the cellular biology and heterogeneity of tissues, we sought an imaging strategy to image multiple markers in 3D in a timely fashion (Supplementary Fig. 1). We first defined a combination of eight fluorophores (out of 21 fluorophores compatible with linear unmixing of lambda stacks[13] that we tested (Fig. 1b and Supplementary Fig. 1b)). Their reference emission spectra were used for accurate unmixing during single-scan acquisition (Supplementary Fig. 2) without the need for individual fluorophore control samples, a major advantage compared to recent methods relying on postacquisition compensation, thereby generating additional data files[1,2,14]. When performing on-the-fly spectral unmixing, equal signal detection is required, which is challenging for eight fluorophores and cannot be achieved through adjusting laser power or detection settings. To overcome this issue, we developed a large-content intensity equalization assay for mLSR-3D-imaging to ensure balanced fluorescent intensities through the immunolabeling process. Using this assay, we tested over 60 antibodies and dyes for optimal eight-color staining (Supplementary Fig. 1 and Supplementary Table 1) and selected five markers of interest based on recent single-cell RNA-sequencing (scRNA-seq) data[15] to label a broad range of early nephrogenic structures of human fetal kidney (HFK) development, as well as 4,6-diamidino-2-phenylindole (DAPI) to stain nuclei, Phalloidin to label the F-actin network and KI67 to mark cycling cells (Fig. 1c–f). To facilitate the use of eight fluorophores, we implemented a 5-day protocol, consisting of three rounds of labeling for flexible use of multiple species of primary antibodies combined with fluorescent secondaries, as well as direct conjugates, followed by a nontoxic clearing step with FUnGI[3] that preserves cell morphology and tissue architecture (Supplementary Fig. 3). This versatile protocol can be applied to a wide range of

[1]Princess Máxima Center for Pediatric Oncology, Utrecht, the Netherlands. [2]Cancer Genomics Netherlands, Oncode Institute, Utrecht, the Netherlands. [3]QVQ Holding BV, Utrecht, the Netherlands. [4]Department of Anatomy and Embryology, Leiden University Medical Center, Leiden, the Netherlands. [5]These authors contributed equally: Ravian L. van Ineveld, Michiel Kleinnijenhuis. ✉e-mail: a.c.rios@prinsesmaximacentrum.nl

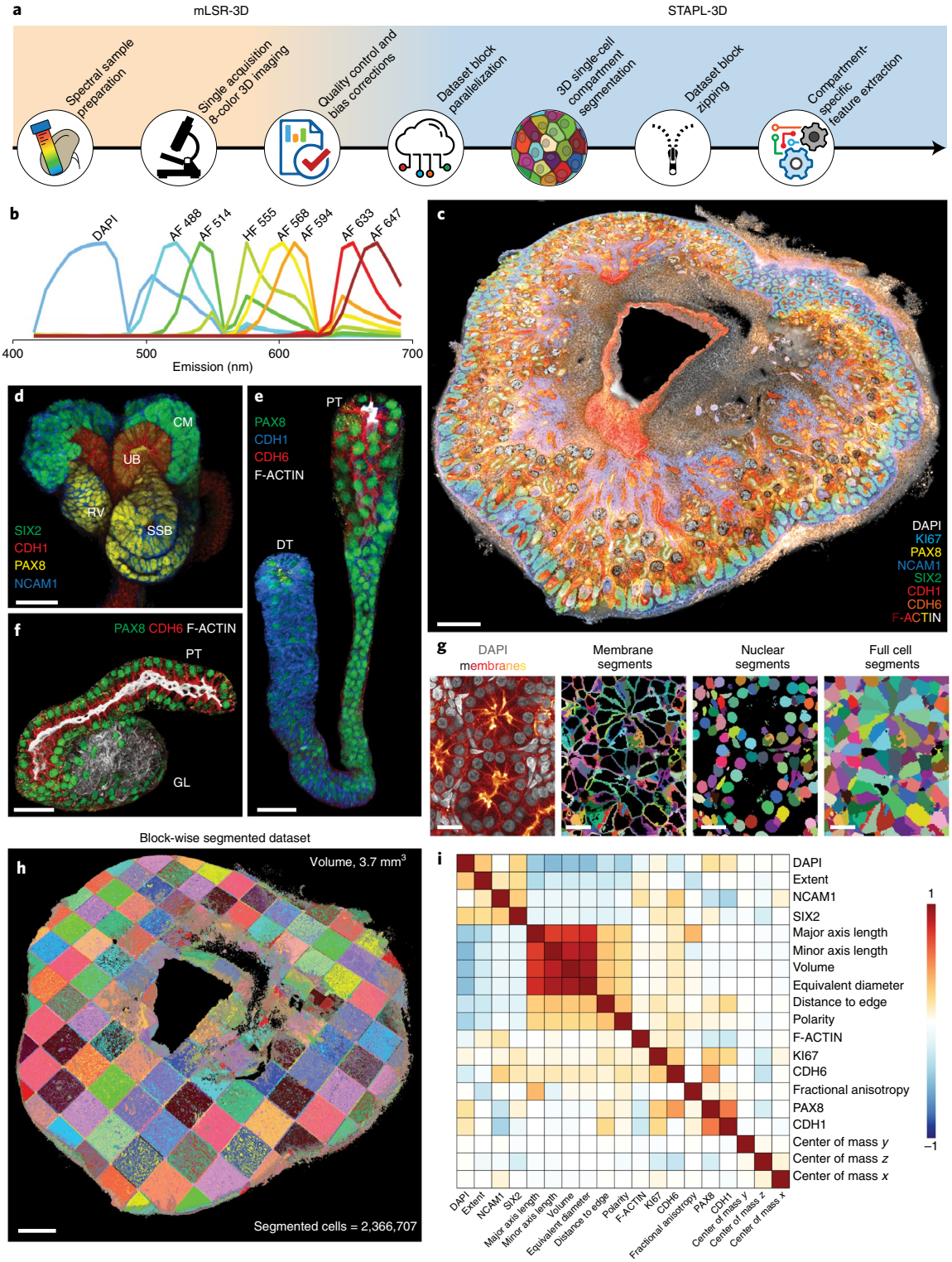


Fig. 1 | mLSR-3D imaging and STAPL-3D. a, Schematic overview of mLSR-3D and STAPL-3D. **b**, Normalized emission reference spectra. **c**, mLSR-3D visualization of HFK (16 weeks of gestation) labeled for DAPI (gray), KI67 (cyan), PAX8 (yellow), NCAM1 (blue), SIX2 (green), CDH1 (red), CDH6 (orange) and F-ACTIN (gradient, red-yellow-white). Scale bar, 500 µm. **d**–**f**, 3D zoomed images of masked nephrogenic structures. **d**, CM, cap mesenchyme; RV, renal vesicle; SSB, S-shaped body; UB, ureteric bud. SIX2 (green), CDH1 (red), PAX8 (yellow) and NCAM1 (blue). **e**, Loop of Henle with proximal tubule (PT) connecting to the distal tubule (DT). PAX8 (green), CDH1 (blue), CDH6 (red) and F-ACTIN (gray). **f**, Proximal tubule (PT) connecting to the glomerulus (GL). PAX8 (green), CDH6 (red) and F-ACTIN (gray). Scale bars, 50 µm. **g**, Optical section demonstrating cell compartment segmentation with STAPL-3D. DAPI (gray) and weighted mean of all membrane channels (red-yellow-white gradient). Segments are randomly colored. Scale bar, 20 µm. **h**, Volumetric rendering of the block-wise segmentation. Number of blocks, 182. Scale bar, 500 µm. These experiments (**c**–**h**) were perfomed independently at least four times with similar results (Supplementary Fig. 10). **i**, Pearson correlation heatmap of selected features, reordered by hierarchical clustering.

tissues, as demonstrated by eight-color mLSR-3D imaging of xenografted human organoid-derived breast tumors (Supplementary Fig. 4a), associated breast cancer organoids cultured in vitro (Supplementary Fig. 4b) and biopsy-derived human central nervous system tumor material (Supplementary Fig. 5). Therefore, this method enables acquisition of large-scale, multi-dimensional 3D datasets with drastic reduction in overall acquisition time, photobleaching by repetitive illumination and data preprocessing and storage requirements.

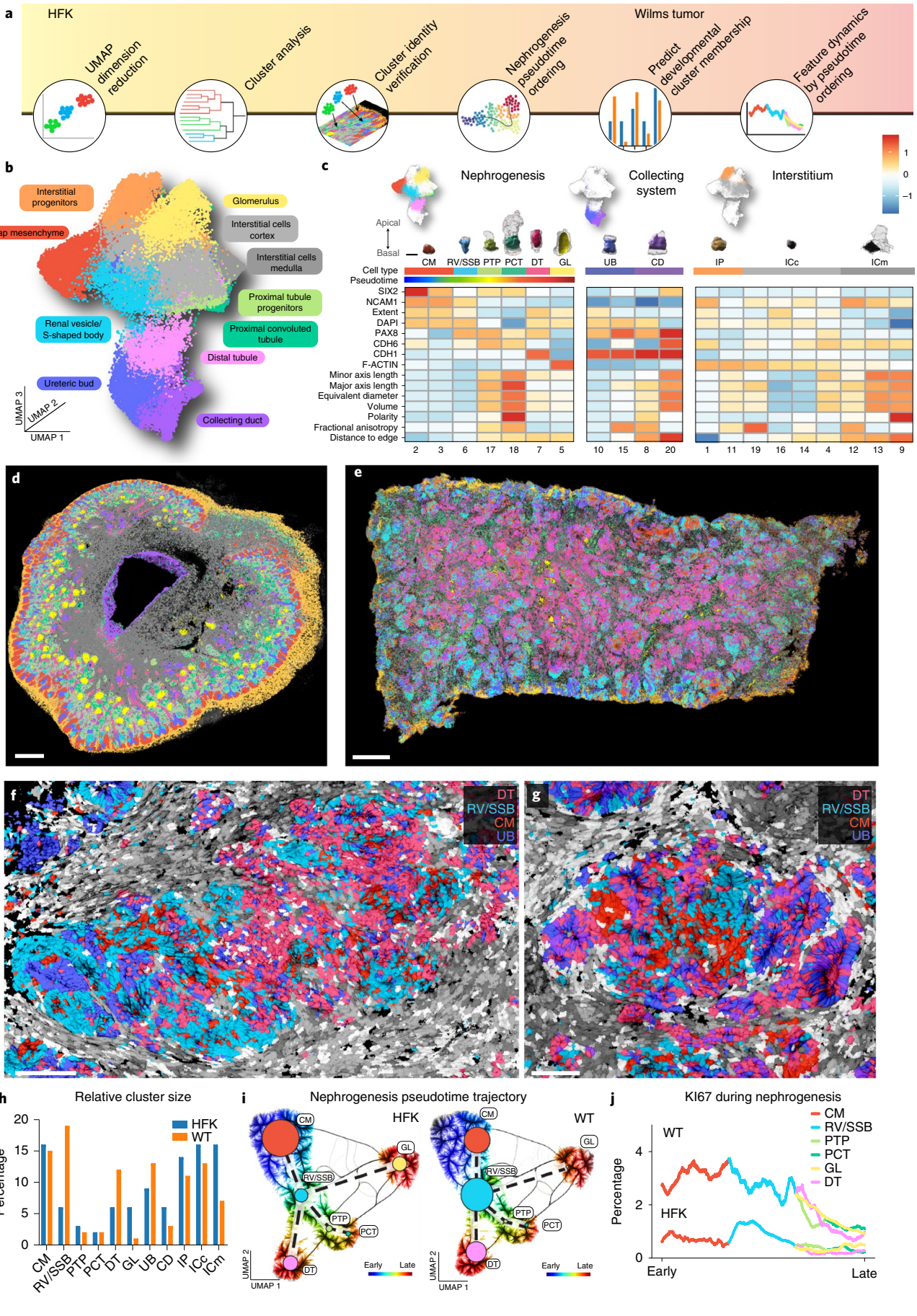
We then developed the STAPL-3D pipeline for single-cell feature extraction from large 3D imaging datasets (Fig. 1a). First, to optimize mLSR-3D datasets for subsequent analysis, we implemented the STAPL-3D preprocessing module (Supplementary Fig. 6a). It includes a new channel-specific shading correction (Supplementary Fig. 6b,d) and a 3D inhomogeneity correction developed for magnetic resonance imaging[16] to reduce technical background variations (Supplementary Fig. 6c,d). Furthermore, for high autofluorescence, observed in the AF488 channel in the kidney, we used machine learning to generate voxelwise probability map[17] for KI67, enabling accurate quantification of cycling cells (Supplementary Fig. 7 and Supplementary Table 2). Next, the STAPL-3D segmentation module (Supplementary Fig. 8a) segments the dataset into individual cells and subdivides each cell into nucleus and membrane (Fig. 1g). STAPL-3D makes optimal use of mLSR-3D data by combining membrane and nucleus channels to generate seeds, followed by a two-step watershed procedure expanding the seed into the nucleus and then filling the cell to the membrane boundary (Supplementary Figs. 8b and 9c). For scalable processing, we designed STAPL-3D to be compatible with high-performance computing for complete segmentation in a couple of hours, by distributing the various analysis steps over volumes, channels and datablocks (Fig. 1h and Supplementary Fig. 10). Yet STAPL-3D also runs efficiently on laboratory workstations. Splitting the dataset into blocks generates seams of partially segmented cells touching the block borders, either resulting in substantial data loss by excluding them[18] or introducing artifacts to these cells. Therefore, we developed a zipping module that identifies erroneous segments, resegments them, and merges the blocks back into a single seamless segmented volume (Supplementary Fig. 8c and 9d).

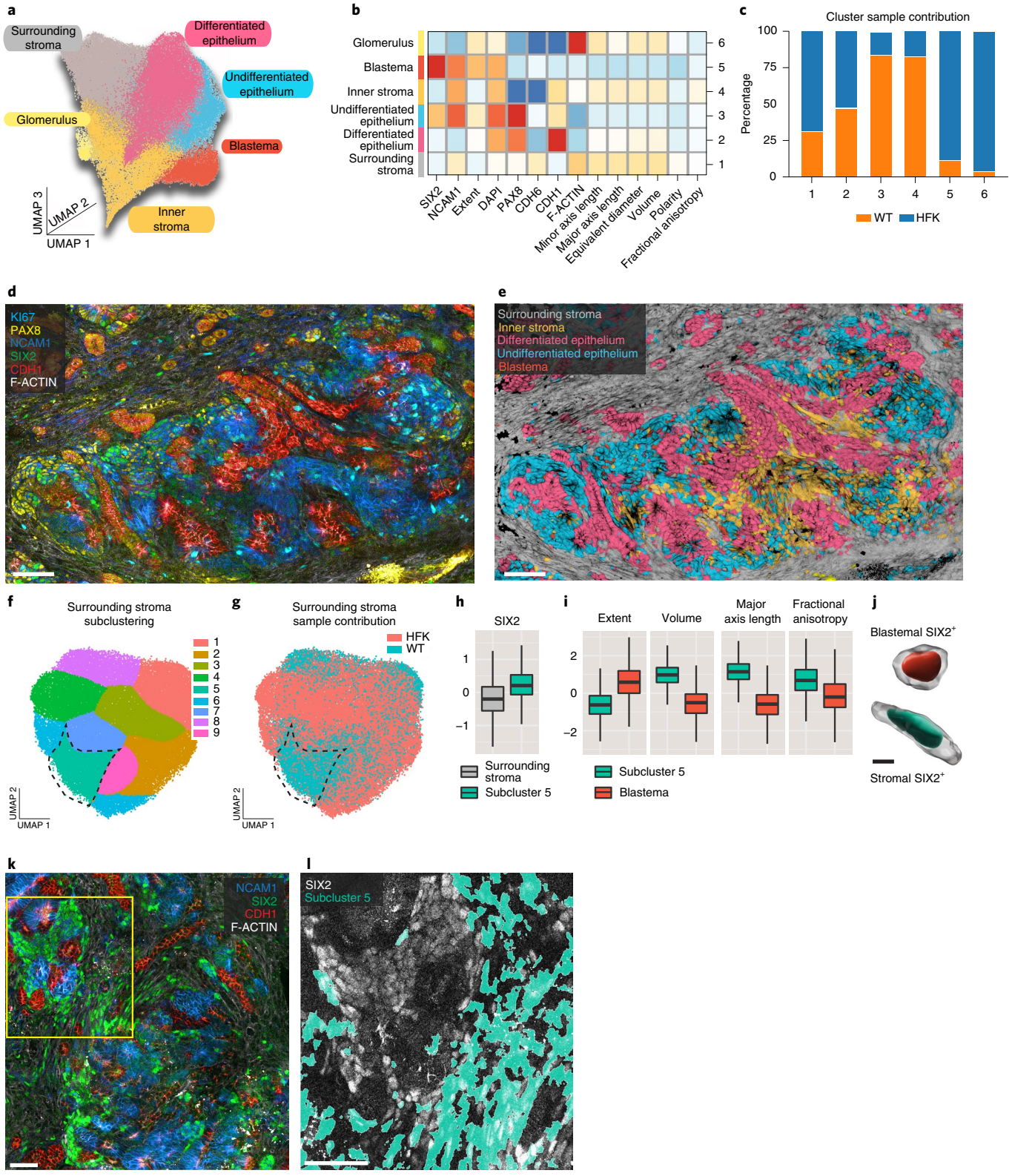
To achieve maximum use of STAPL-3D, we provide the option to use state-of-the-art deep learning segmentation methods within the pipeline (STAPL-3D^{DL}), by integrating a 3D universal network (3D-UNET)[10] to predict membrane probability and StarDist[11] to predict individual nuclei (Supplementary Fig. 11). Because manual segmentation proved practically unfeasible for mLSR-3D datasets (with 80 h of labor required for 569 cells, which was insufficient for model training), we also provide a STAPL-3D module to generate large training datasets (Supplementary Fig. 11a) by coacquisition of mLSR-3D data at typical resolution (yielding the training data) and at very high resolution (yielding the training labels at the same location). Furthermore, using these labeled datasets, we offer a segmentation parameter tuning module that uses Bayesian optimization to automatically choose parameters that result in the best segmentation quality (Supplementary Table 3, STAPL-3D^{FT}). Segmentation accuracy of the modules was assessed by comparison to an extensive ($n = 14{,}717$ cells) and a diverse set of cells (from ten different areas of the kidney) that was segmented with high fidelity from datasets coacquired at high resolution and manually curated afterward to yield a ground truth dataset. Dice overlap, precision, recall and F scores were computed as accuracy metrics (Supplementary Fig. 12 and Supplementary Table 3). We obtained the highest $F_{1.5}$-score of 0.81 ($\pm$0.012 s.e.m.) for STAPL-3D^{DL} trained on coacquired mLSR-3D datasets (Supplementary Fig. 12b) followed by STAPL-3D^{FT} ($F_{1.5} = 0.75 \pm 0.014$), both demonstrating a significant increase in performance over the generic deep learning model ($F_{1.5} = 0.71 \pm 0.014$) and the nontuned STAPL-3D pipeline ($F_{1.5} = 0.72 \pm 0.017$). Furthermore, comparing extracted morphological features with the ground truth, showed little morphological divergence. Overall average percentual increases and decreases with respect to the ground truth were 5.929% ($\pm$2.905) and 7.198% ($\pm$1.740) for STAPL-3D^{DL} trained and 7.821% ($\pm$2.768) and 8.912% ($\pm$2.220) for STAPL-3D^{FT} (Supplementary Fig. 12c). This analysis thus confirms that once trained on the appropriate data, STAPL-3D^{DL} increases segmentation accuracy. Nevertheless, STAPL-3D already offers a robust segmentation pipeline with a good performance.

STAPL-3D extracts molecular marker intensities, as well as spatial and 3D morphological properties per segmented cellular compartment. By default, features are computed for the cell and the membranal and nuclear subsegments. Moreover, we show that mLSR-3D and STAPL-3D pipelines can be adapted to extract features from the cytosolic and even mitochondrial compartment when using Airyscan 3D imaging (Supplementary Fig. 13). The division into cellular compartments can be exploited to define compound features, for example cell polarity, estimated from the centers of mass of a cell and nucleus (Supplementary Fig. 8d). We can obtain a complete set of approximately 800 features extracted from full cell, nucleus and membrane segments (Supplementary Table 4), with the option to select the features most relevant for the particular application (Fig. 1i). Altogether, STAPL-3D offers a scalable, modular and tunable analysis framework for advanced image preprocessing and cellular compartment-specific segmentation, toward reliable 3D feature extraction and profiling of millions of cells within tissue.

To showcase the potential of our mLSR-3D and STAPL-3D framework, we next performed spatio-phenotypic patterning of Wilms tumor—a pediatric kidney cancer (Fig. 2 and Supplementary Fig. 14). Prevalence of Wilms tumor in early childhood has been related to corruption of fetal nephrogenesis and, indeed, these tumors present with aberrant fetal cells[19,20]. Therefore, we aim to elucidate the in situ developmental patterning of Wilms tumor in relation to HFK. To create a spatio-phenotypic reference map, we defined 11 known cell populations in our HFK sample (gestational week 16),

Fig. 2 | Spatio-phenotypic patterning of HFK and Wilms tumor reveals expanded, cycling, early epithelial compartment in Wilms tumor. a, Schematic overview of the analysis strategy. **b**, 3D UMAP rendered for 50,000 HFK cells. Colors and labels correspond to cell identity. **c**, Heatmap of log-scaled median feature values (blue-white-red gradient) per identified cluster, subdivided into components of the HFK: nephrogenesis (CM, cap mesenchyme; RV/SSB, renal vesicle/S-shaped body; PTP, proximal tubule progenitor; PCT, proximal convoluted tubule; DT, distal tubule; GL, glomerulus), collecting system (UB, ureteric bud; CD, collecting duct) and interstitium (IP, interstitial progenitors; ICc, interstitial cells cortex; ICm, interstitial cells medulla). Clusters are numbered from 1 to 20 according to descending cluster size. Typical 3D segmented examples of each cell type are displayed above the heatmap, oriented apical to basal. Scale bar, 10 µm. **d**,**e**, 3D backprojection of the HFK (**d**) and Wilms tumor (**e**) showing all single-cell segments colored for cell type identity. Scale bars, 500 µm. **f** Optical section showing backprojected cell types of a Wilms tumor region consisting predominantly of distal tubule, RV/SSB and ureteric bud. Scale bar, 70 µm. **g**, Optical section showing backprojected cell types of a Wilms tumor region containing a cell cluster of more undifferentiated cap mesenchyme. Scale bar, 70 µm. **h**, Bar graph depicting relative cluster sizes per dataset in percentages. **i**, UMAP depicting pseudotime ordering of HFK (left panel) and Wilms tumor (WT, right panel) cells belonging to nephrogenic clusters. Circle sizes correspond to the number of cells within each cluster and pseudotime is depicted by the rainbow gradient (early, blue to red, late). **j**, KI67-positive fraction for Wilms tumor (top) and HFK (bottom) plotted along the pseudotime trajectory of nephrogenic development. Line colors correspond to cell identities.


a
HFK
UMAP dimension reduction
Cluster analysis
Cluster identity verification
Nephrogenesis pseudotime ordering
Wilms tumor
Predict developmental cluster membership
Feature dynamics by pseudotime ordering
b
Interstitial progenitors
Glomerulus
Cap mesenchyme
Interstitial cells cortex
Interstitial cells medulla
Proximal tubule progenitors
Renal vesicle/ S-shaped body
Proximal convoluted tubule
Distal tubule
Ureteric bud
Collecting duct
UMAP 3
UMAP 2
UMAP 1
c
Nephrogenesis
Collecting system
Interstitium
Apical
Basal
CM RV/SSB PTP PCT DT GL UB CD IP ICc ICm
Cell type
Pseudotime
SIX2
NCAM1
Extent
DAPI
PAX8
CDH6
CDH1
F-ACTIN
Minor axis length
Major axis length
Equivalent diameter
Volume
Polarity
Fractional anisotropy
Distance to edge
2 3 6 17 18 7 5
10 15 8 20
1 11 19 16 14 4 12 13 9
d
e
f
DT
RV/SSB
CM
UB
g
DT
RV/SSB
CM
UB
h
Relative cluster size
Percentage
HFK
WT
CM RV/SSB PTP PCT DT GL UB CD IP ICc ICm
i
Nephrogenesis pseudotime trajectory
HFK
WT
CM
GL
RV/SSB
PTP
PCT
DT
Early
Late
UMAP 2
UMAP 1
j
KI67 during nephrogenesis
WT
HFK
Percentage
CM
RV/SSB
PTP
PCT
GL
DT
Early
Late



distributed over three components of the developing kidney: nephrogenesis, collecting system and interstitium[15]. This was achieved through 3D uniform manifold approximation and projection (UMAP)[21] and clustering of the 2.1 million cells × 19 features data matrix (Fig. 2b and Supplementary Fig. 15c), assigning the resulting 20 clusters to a particular population, based on molecular markers, but also indispensably aided by morphological features and spatial location (Fig. 2c,d and Supplementary Figs. 15 and 16a,b). Having captured the spatio-phenotypic single-cell landscape of HFK in a classifier, cell types could also be predicted for the 1.8 million cells segmented from the Wilms tumor sample. Backprojection of population identity into the dataset revealed a highly disorganized spatial pattern in Wilms tumor compared to HFK (Fig. 2d,e), yet nephrogenic-like structures could be identified. These structures

Fig. 3 | Comparative analysis reveals spindle-shaped SIX2⁺ cells in surrounding tumor stoma. a, Joint 3D UMAP rendered for 50,000 HFK and Wilms tumor cells. Colors and labels correspond to cluster identity. **b**, Heatmap of log-scaled median feature values (blue-white-red gradient) per identified cluster: clusters are numbered 1–6 according to descending cluster size. **c**, Bar graph depicting relative contribution of each sample per cluster in percentages. **d**,**e**, Optical sections showing fluorescent markers KI67 (cyan), PAX8 (yellow), NCAM1 (blue), SIX2 (green), CDH1 (red) and F-ACTIN (gray) (**d**) and backprojected cluster identities (**e**) of a representative Wilms tumor (WT) region. Scale bars, 70 µm. **f**,**g**, UMAP rendered for 50,000 surrounding stroma cells. Colors and labels correspond to subcluster identity (**f**) or sample identity (**g**): blue, Wilms tumor; red, HFK. **h**, Boxplots showing log-scaled median SIX2 value of subcluster 5 (n = 194,218 cells) compared to the surrounding stroma main cluster (n = 1,650,651 cells). Center, median; bounds, Q1–Q3, whiskers extend to minimum/maximum limited to 1.5 times the IQR. **i**, Boxplots showing log-scaled median values for extent, volume, major axis length and fractional anisotropy of subcluster 5 (n = 194,218 cells) compared to the blastema main cluster (n = 342,579 cells). Center, median; bounds, Q1–Q3, whiskers extend to minimum/maximum limited to 1.5 times the IQR. **j**, Typical 3D segmented examples of a SIX2⁺ blastemal (left) and SIX2⁺ stromal cell (right). Scale bar, 10 µm. **k**, Optical section showing fluorescent markers NCAM1 (blue), SIX2 (green), CDH1 (red), F-ACTIN (gray) of a representative Wilms tumor region. Scale bar, 50 µm. **l**, Magnification of indicated yellow area with backprojected subcluster 5 identities and SIX2 (gray). Scale bar, 50 µm.

consisted predominantly of distal tubule, ureteric bud and renal vesicle/S-shaped body (RV/SSB) cells (Fig. 2f), and sporadically contained cell clusters of more undifferentiated cap mesenchyme (CM)-like cells (Fig. 2g). Indeed, in relation to HFK, the epithelial components of this particular Wilms tumor were enlarged: 1.2 times for ureteric bud, 1.6 times for distal tubule and, most notably, 2.8 times for early epithelial RV/SSB (Fig. 2h). Pseudotime ordering of nephrogenic cells revealed a single trajectory for early progenitors (CM) to committed progenitors (RV/SSB), branching into three late populations (distal tubule (DT), proximal tubule progenitor/proximal convoluted tubule and glomerulus (GL)) (Fig. 2i). The position of the center of mass in this pseudotime UMAP showed a shift for the CM node toward the enlarged RV/SSB node in Wilms tumor, indicative of a more committed progenitor fate for this cluster as compared to HFK. Inclusion of KI67 in our set of markers allowed us to provide insight into the mechanism underlying this developmental pattern (Fig. 2j). From the pseudotime-ordered cells, we could identify a peak of cycling cells during the RV/SSB stage in HFK. Wilms tumor showed overall increased cycling compared to HFK; in particular high in RV/SSB, but also in their developmental progenitors (CM) (Fig. 2j). Thus, the enlarged RV/SSB cluster in the Wilms tumor sample (Fig. 2h) likely results from both intrinsic cycling properties of this cluster, as well as a transitioning progenitor population from the CM fueling this compartment. Hence, through profiling population distribution and pseudotime ordering, we could begin to untangle the in situ heterogeneity of Wilms tumor in relation to its developmental origin.

To dive deeper into the spatio-phenotypic traits specific to the Wilms tumor, we next created a joint UMAP for HFK and Wilms tumor and identified six cellular clusters that largely reflect conventional Wilms tumor classification (epithelium, stroma and blastema[22]) (Fig. 3a,b). Based on their spatio-phenotypic features, we describe them as differentiated and undifferentiated epithelium, blastema, two stroma clusters, but also a small glomerulus population (Fig. 3b and Supplementary Fig. 15). Although varying in contribution, all six clusters contained both healthy kidney and tumor cells, confirming the strong fetal resemblance of this tumor (Fig. 3c). Cluster backprojection reveals one stroma-like compartment surrounding the epithelial/blastemal clusters (the surrounding stroma) while the other locates within these structures (the inner stroma) (Fig. 3d,e). This nicely demonstrates the usefulness of maintaining tissue-context to reveal differential spatial embedding of populations. The relative cluster sizes resulting from our classification (46.7% epithelium, 51% stroma and 2.3% blastema) (Supplementary Table 5), closely agree with histopathological scoring of the Wilms tumor sample (50% epithelium, 45% stroma and 5% blastema), providing confidence in the obtained classification. Yet our approach goes beyond conventional classification and offers a more in-depth characterization (that is, differentiated versus undifferentiated epithelium and two spatially resolved stromal-like compartments). In addition, we identified a tumor-specific population within the surrounding stroma through subclustering (Fig. 3f,g). Cells belonging to subcluster 5 showed a high expression of SIX2 compared to the remaining surrounding stroma (Fig. 3h), and are spindle-shaped (Fig. 3i), unlike conventional SIX2⁺ round-shaped blastemal cells in Wilms tumor (Fig. 3i,j, high-extent). This may be of particular significance, because SIX2 is involved in maintaining the undifferentiated and proliferative state of HFK CM and Wilms tumor blastema cells[23], the latter known to be associated with poor prognosis if prevalent after chemotherapy[22]. Twenty-four percent of Wilms tumor cells of this subcluster 5 expressed SIX2 to a similar intensity as blastema cells (that is, falling within or above the blastema interquartile range). Even though they represent only 2% of the entire Wilms tumor sample, these SIX2-high stromal-like cells are present in substantial, and, perhaps, thereby consequential amounts (36,242 cells). Although clinical importance of the identified cell profiles remains to be determined, we demonstrated that the combined application of mLSR-3D and STAPL-3D offers the potential to generate new insights into tumor biology by accurate cell subset quantification and identification of new spatio-phenotypic signatures.

Discussion

In sum, we here provide a targeted in situ profiling approach to exploit molecular, morphological and spatial features of millions of cells from 3D imaging data. In line with recent advances in multiplexed proteomics and spatial transcriptomics[1,3,4,24–26], we envision our single-cell technology a key step forward toward unraveling the complex cellular organization of organs and their associated tumors with particular promise for capturing essential spatio-phenotypic hallmarks of tumorigenesis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00926-3.

Received: 11 June 2020; Accepted: 16 April 2021;
Published online: 3 June 2021

References

1. Coutu, D. L., Kokkaliaris, K. D., Kunz, L. & Schroeder, T. Multicolor quantitative confocal imaging cytometry. *Nat. Methods* **15**, 39–46 (2018).
2. Li, W., Germain, R. N. & Gerner, M. Y. High-dimensional cell-level analysis of tissues with Ce3D multiplex volume imaging. *Nat. Protoc.* **14**, 1708–1733 (2019).
3. Rios, A. C. et al. Intraclonal plasticity in mammary tumors revealed through large-scale single-cell resolution 3D imaging. *Cancer Cell* **35**, 618–632.e6 (2019).
4. Segovia-Miranda, F. et al. Three-dimensional spatially resolved geometrical and functional models of human liver tissue reveal new aspects of NAFLD progression. *Nat. Med.* **25**, 1885–1893 (2019).

5. Steinert, E. M. et al. Quantifying memory CD8 T cells reveals regionalization of immunosurveillance. *Cell* **161**, 737–749 (2015).
6. Mosaliganti, K. R., Noche, R. R., Xiong, F., Swinburne, I. A. & Megason, S. G. ACME: automated cell morphology extractor for comprehensive reconstruction of cell membranes. *PLoS Comput. Biol.* **8**, e1002780 (2012).
7. Stegmaier, J. et al. Real-time three-dimensional cell segmentation in large-scale microscopy data of developing embryos. *Dev. Cell* **36**, 225–240 (2016).
8. McQuin, C. et al. CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).
9. Dunn, K. W. et al. DeepSynth: Three-dimensional nuclear segmentation of biological images using neural networks trained with synthetic data. *Sci. Rep.* **9**, 18295 (2019).
10. Wolny, A. et al. Accurate and versatile 3D segmentation of plant tissues at cellular resolution. *Elife* **9**, 1–34 (2020).
11. Weigert, M., Schmidt, U., Haase, R., Sugawara, K. & Myers, G. Star-convex polyhedra for 3D object detection and segmentation in microscopy. in Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020 3655–3662 (2020). https://doi.org/10.1109/WACV45572.2020.9093435
12. Zhao, S. et al. Cellular and molecular probing of intact human organs. *Cell* **180**, 796–812.e19 (2020).
13. Kraus, B., Ziegler, M. & Wolff, H. Linear fluorescence unmixing in cell biological research. *Mod. Res. Educ. Top. Microsc.* **2**, 863–872 (2007).
14. Valm, A. M. et al. Applying systems-level spectral imaging and analysis to reveal the organelle interactome. *Nature* **546**, 162–167 (2017).
15. Hochane, M. et al. Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLoS Biol.* **17**, e3000152 (2019).
16. Tustison, N. J. et al. N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
17. Berg, S. et al. Ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods* **16**, 1226–1232 (2019).
18. Gut, G., Herrmann, M. D. & Pelkmans, L. Multiplexed protein maps link subcellular organization to cellular states. *Science* **80**, 361 (2018).
19. Young, M. D. et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594–599 (2018).
20. Young, M. D. et al. Single cell derived mRNA signals across human kidney tumors. Preprint at *bioRxiv* https://doi.org/10.1101/2020.03.19.998815 (2020).
21. McInnes, L. PCA, t-SNE, and UMAP: modern approaches to dimension reduction. PyData Conference 2018 (2018).
22. Reinhard, H. et al. Outcome of relapses of nephroblastoma in patients registered in the SIOP/GPOH trials and studies. *Oncol. Rep.* **20**, 463–467 (2008).
23. Wegert, J. et al. Mutations in the SIX1/2 pathway and the DROSHA/DGCR8 miRNA microprocessor complex underlie high-risk blastemal type Wilms tumors. *Cancer Cell* **27**, 298–311 (2015).
24. Glaser, A. K. et al. Multi-immersion open-top light-sheet microscope for high-throughput imaging of cleared tissues. *Nat. Commun.* **10**, 2781 (2019).
25. Stoltzfus, C. R. et al. CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. *Cell Rep.* **31**, 107523 (2020).
26. Merritt, C. R. et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat. Biotechnol.* **38**, 586–599 (2020).

Methods

Ethics statement. The collection and use of HFK were approved by the Medical Ethics Committee from the Leiden University Medical Center (P08.087). The gestational age was determined by ultrasonography, and the tissue was obtained from women undergoing elective abortion. The material was donated with written informed consent. Human organoid samples were retrieved from a biobank through the Hubrecht Organoid Technology (HUB, www.hub4organoids.nl). Wilms tumor and pediatric spinal ependymoma biopsy material was obtained from the biobank of the Princess Máxima Center (PMCLAB2019.037). Authorizations were obtained from the medical ethical committee of the UMC Utrecht (METC UMCU) to ensure compliance with the Dutch 'medical research involving human subjects' act and informed consent was obtained from donors where appropriate. Children with Wilms tumor receive neoadjuvant cytotoxic treatment before nephrectomy, as per Dutch practice. Breast tumor tissue was obtained from xenograft mouse models approved by the Animal Welfare Committee of the Princess Máxima Center and established in compliance with both local and international regulations.

Tissue fixation and blocking. Dissected HFK, xenografted breast tumors and human spinal ependymoma tissue were immersed in 5 ml of 4% paraformaldehyde at pH 7.4 overnight on ice. After fixation, samples were washed in PBT (1 ml of Tween-20 in 1 l of phosphate-buffered saline (PBS)) for 15 min and then blocked in 5–10 ml of wash buffer 1 (2 ml of Tween-20, 2 ml of Triton X-100, 2 ml of 10% SDS, 2 g of bovine serum albumin (BSA) in 1 l of PBS) for 2–5 h depending on the size of the sample. Breast tumor organoids were harvested and fixed, as previously described[27,28].

Immunolabeling. We performed eight-color immunolabeling using off-the-shelf antibodies (Supplementary Table 1) in a three-round staining protocol (Supplementary Fig. 3). In the first round, the tissue was incubated with nonlabeled primary antibodies of different species. Second, we used fluorescently labeled secondary antibodies targeted against the different species. To overcome the limited diversity in antibody species, we made use of direct fluorescently labeled primary antibodies and dyes in a third labeling round. Washing and incubation steps were performed in wash buffer 2 (1 ml Triton X-100, 2 ml of 10% SDS, 2 g of BSA in 1 l of PBS).

Tissue clearing. Samples were optically cleared by three stepwise incubations (1 h at room temperature) of increasing concentration (25/50/75%) of FUnGI clearing agent diluted in PBS. The final incubation step with 100% FUnGI was performed overnight at 4 °C.

Spectral library acquisition. Pieces of HFK tissue were labeled with single fluorophores and lambda stack images were taken for every fluorophore, with an excitation filter combination of main beam splitter (MBS) 405 and MBS 488/561/633. The acquired images were unmixed with the auto find function in the Zeiss Zen Black software and the obtained reference signature spectra (Supplementary Fig. 1b) were saved for using the online fingerprinting mode.

Spectral imaging with linear unmixing and quantitative measurement of the spectral unmixing. Imaging was performed on a Zeiss LSM880 with a 32-channel spectral detector using a ×25 0.8 numerical aperture (NA) multi-immersion objective with a working distance of 500 µm. The online fingerprinting mode was used to separate the eight fluorophores into distinct channels during acquisition. Tile scans were acquired with 10% overlap of *z*-stacks and with a pixel dwell time of 2 µs and a voxel size 0.33 × 0.33 × 1.2 µm. Data were digitized with 16 bits per voxel. With these acquisition settings, on average 4 mm^3 of tissue was imaged, requiring a scan time of around 16 h. Online fingerprinting removes the need to store the 32-channel lambda data. Therefore, the unmixed eight-channel data has a fourfold reduction in size compared to 32-channel lambda data, expediting downstream analysis.

Linear unmixing accuracy was assessed in the Zeiss Zen Blue software by processing of a raw eight-channel mLSR-3D lambda stack (Supplementary Fig. 2). The linear unmixing processing tool was used with the option to display channels with statistical confidence. The statistical confidence considers the noise from the channels, the bandwidth and position and the quality of the reference spectra. The resulting images visualize the statistical confidence for every pixel by a percentage-based fire color map for interpretation.

mLSR-3D equalization assay. When performing on-the-fly spectral unmixing imaging, multiple fluorophores are excited by the same laser and all fluorophores are detected by the same detector. In this setup, large fluorescence intensity differences cannot be compensated by adjusting laser power or detection settings. To tackle this challenge, we developed the mLSR-3D equalization assay (Supplementary Fig. 1a).

Sample preparation. Tissue samples were manually cut to approximately 1 mm × 1 mm × 1 cm and embedded in 4% UltraPure Low Melting Point Agarose (Thermo Fisher) in PBS in a 96-well flat bottom plate. Sections of 500 µm were cut with a vibratome (Leica VT 1200S) set to a velocity of 0.60 mm s^{-1}, an amplitude of 3.00 mm and an angle of 18°. To allow simultaneous sample preparation for single stains of over 60 antibodies and dyes, agarose embedded tissue sections were placed into a 96-well filter plate (Thermo Fisher), sealed at the bottom and single stains were performed according to the mLSR-3D immunolabeling protocol with the following adaptations: for washing and blocking, 200 µl of buffer was dispensed to each well with a Multipette pipette (Eppendorf). Antibody dilutions were prepared in eight-tube PCR strips and dispensed with a multichannel pipette (Gilson). For incubation steps, the 96-well filter plates were sealed and placed inversely inside a dark box on an orbital platform mixer at 4 °C. To discard buffer, the bottom seal was removed and the plate was attached to a waste container and centrifuged for 5–10 min at 500*g*, until no remaining buffer or bubbles were present. For clearing and subsequent imaging, samples were transferred to a 96-well imaging plate prefilled with FUnGI.

Data acquisition. To calibrate laser power, a laser power meter (FieldMate, 1098297, Coherent) was connected to the ×25 objective. The percentage of each laser was adjusted individually to an output power of 19–21, 10.3–10.5, 26–27.5 and 104–108 µW for the 405, 488, 561 and 633 nm excitation lasers, respectively. For linear unmixing, standard acquisition settings of the online fingerprinting mode were used, as described above.

Signal intensity quantification and normalization. For calculation of signal intensities, we designed an automated quantification procedure. The metric used is the contrast-to-noise ratio of the structures expressing the markers (signal-of-interest) with respect to the tissue background signal. This was calculated from a three-component segmentation (signal-of-interest, background and noise regions). For each optical section, data were smoothed with a Gaussian filter with $\sigma = 20$ µm and thresholded to generate masks for tissue and nontissue, where the optimal thresholds were determined in ITK-SNAP[29] for each optical section individually. Voxels clipping at the upper end of the 16-bit scale were removed from the tissue mask. The tissue mask was divided in the signal-of-interest mask—defined as the set of voxels in the 99th percentile within the tissue mask: only retaining clusters of >3 connected voxels—and the background mask—that is, all other voxels in the tissue mask. The contrast-to-noise metric was calculated as the difference between the median of the values in the signal-of-interest mask and the median of the values in the background mask, divided by the standard deviation of the values in the nontissue mask. To classify this signal as high, medium or low for secondary antibodies, two equidistant thresholds were set between zero and the highest value obtained to create three intensity categories. With the guidance of this classification, combinations of primary and secondary antibodies were optimized to achieve an equal intensity for each marker excited by the same laser (Supplementary Fig. 1c–e).

Multi-resolution coacquisition 3D imaging. To generate the extensive ground truth datasets used for deep learning training, accuracy assessment and fine tuning, matching 3D *z*-stacks at typical and high resolution from the same position were acquired with a ×25 0.8 NA multi-immersion objective (working distance 500 µm) at voxel size 0.33 × 0.33 × 1.2 $µm^3$ and with a ×63 1.4 oil-immersion objective (working distance 140 µm) at voxel size 0.13 × 0.13 × 0.39 µm. Standard acquisition settings of the online fingerprinting mode were used for both resolution types, as described above. To reduce the effects of photobleaching, the typical resolution was acquired first before acquisition of the high-resolution data.

Airyscan super-resolution 3D imaging. Imaging was performed on a Zeiss LSM880 with an Airyscan detector using a ×25 0.8 NA multi-immersion objective with a working distance of 500 µm and ×63 1.4 NA oil-immersion objective with a working distance of 140 µm. *Z*-stacks were acquired with a pixel dwell time of 2.05 µs (×25 objective) and 4.10 µs (×63 objective) with a voxel size of 0.185 × 0.185 × 0.576 $µm^3$ (×25 objective) and 0.0628 × 0.0628 × 0.187 $µm^3$ (×63 objective). Data were digitized with 16 bits per voxel.

Shading correction. *Z*-stack shading was corrected in individual channels by estimating intensity profiles over both the *x* and *y* dimensions (Supplementary Fig. 6b). Concatenating all *z*-stacks and masking any value higher than the chosen noise threshold ($I > 1{,}000$), the median values over *x* (or *y*) were computed for each *yz* (or *xz*) coordinate (Supplementary Fig. 6b, first panel). The illumination profile over *x* (or *y*) was then estimated by averaging over planes. Because empty planes do not provide reliable estimates, a subset of planes was selected by taking the median over *y* (or *x*), retaining only the planes with median values larger than the 0.80 quantile (second panel, colored traces). A third-order polynomial was fit to the intensity profile (second panel, black dashed trace) and normalized to the highest value in the fitted profile (second panel, black solid trace). These fitted *x* and *y* profiles were then multiplied to generate an estimated bright image (third panel), one for each channel in the dataset. Shading artifacts were then corrected in the Zeiss Zen software (Blue Edition v.2.6) through division of each plane in each *z*-stack with the estimated bright image (fourth panel).

Stitching. *Z*-stacks were stitched in Zeiss Zen (Blue Edition v.2.6) using the central *z*-plane of the DAPI channel (parameters Edge Detector=off, Minimal

Overlap=5%, Maximal Shift=2%, Comparer=Best, Global Optimizer=Best). The estimated *xy* translations for each *z*-stack were then applied to each plane in each channel. Tile-fusion was used (Fuse Tiles=on), but the built-in shading correction was disabled (Correct Shading=off), as it was replaced by our in-house algorithm.

3D inhomogeneity correction. Further inhomogeneity correction was used to correct for low-frequency intensity variation that does not represent specific staining in the stitched dataset (Supplementary Fig. 6c). These variations have multiple sources, including attenuation over depth from the light path as well as effects of clearing agent and antibody penetration during the sample preparation. A low-resolution dataset was generated by downsampling the data in-plane to a voxel spacing of $zyx = 1.2 \times 21 \times 21\,\mu m^3$ (that is, by a factor of 64 for our datasets). The inhomogeneity was estimated for each channel separately using the N4 algorithm[16], a bias field correction algorithm developed for magnetic resonance imaging and implemented in the Insight Toolkit. First, a dataset mask was created by averaging over all channels, smoothing the averaged image ($\sigma = 48\,\mu m$) and thresholding the smoothed image ($I > 1{,}000$). The low-frequency variation was estimated from the data in this mask using 50 iterations at four fitting levels and with five control points for the *b*-spline fit in each dimension. Each channel was corrected by dividing the intensities in the image at the native resolution by the point estimation (trilinear interpolation) of the 3D inhomogeneity field at these locations.

KI67 quantification. For the purpose of quantifying cycling cell populations, a machine-learning-based binary classification of KI67 cells was performed, assigning a positive or negative label to each segment (Supplementary Fig. 7). To robustly reduce the effects of background autofluorescence, a probability map for specific KI67 staining was generated in ilastik (v.1.3.2)[17] through the Pixel Probability workflow. In a datablock of $106 \times 1{,}408 \times 1{,}408$ voxels from the 16-week HFK, all KI67-positive cells were manually annotated with a 3D brush of roughly 20 px in ITK-SNAP (v.3.8.0)[29]. In ilastik, a background class was added by interactive annotation. The classifier was trained with 11 features (Supplementary Table 2). The probability for each voxel to belong to a KI67-positive cell was then predicted for all voxels in HFK and Wilms tumor datasets. KI67-positive voxels were defined by thresholding the KI67 probability ($P > 0.75$). The overlap of this binary mask of KI67-positive voxels was calculated for each segmented nucleus. Cells were marked KI67-positive if the nucleus contained at least five KI67-positive voxels that made up more than 10% of the nucleus volume.

We manually annotated the KI67-positive cells in two additional datablocks to compare the accuracy of the classification with the accuracy of the default approach of thresholding the median KI67 signal intensity over the segment. The manually drawn mask was processed identically to the thresholded KI67 probability prediction (that is, >5 voxels in the segment and >10% overlap with the nucleus) to generate the labels to which the predictions are compared. Evaluating the accuracy metrics (Supplementary Fig. 7k–n), revealed that the optimal threshold for the KI67 probability was correctly picked at $P = 0.75$.

Seed-detection segmentation. The seed-detection segmentation strategy, in brief, consists of the following steps. We generate a mask of membranes and a mask of nuclei and combine them to obtain a mask where the nuclei are separated as much as possible. A distance transform on this combined mask then results in an image that is expected to have peaks in the nuclei centers. These peaks are detected and used as seeds in a watershed algorithm.

The four membrane channels were averaged (weights $w_{NCAM1} = 0.5$; $w_{CADH1} = 0.5$; $w_{CADH6} = 1.0$; $w_{FACTIN} = 1.0$) to obtain a membrane image with optimal coverage and signal-to-noise (Supplementary Fig. 8b, panel 1). Membrane enhancement was performed on this image using the automated cell morphology extractor (ACME) method[6] that enhances planar structure in 3D data (Supplementary Fig. 8b, panel 2). ACME parameters were set to membrane radius, 0.5 µm and neighborhood radius, 1.1 µm. The planarity measure that results of this procedure was thresholded ($I > 0.0005$) to create a membrane mask (Supplementary Fig. 8b, panel 3).

A mask covering the nuclei was generated from the DAPI channel (Supplementary Fig. 8b, panel 4). The DAPI channel was first preprocessed ($proc_{DAPI}$) by gentle greyscale opening and in-plane Gaussian smoothing with $\sigma = 0.33\,\mu m$ (Supplementary Fig. 8b, panel 5). We made use of data-adaptive thresholding to recover nuclei in areas where the DAPI channel exhibited residual dimness in the deepest planes. The Sauvola method[30] was used, setting the parameters window size to [19, 75, 75] and *k* to 0.2, while *r* was kept at the default (half the datatype range). Because the data-adaptive thresholding finds a very low threshold in areas where no nuclei are present, an absolute minimum intensity threshold ($I > 2{,}000$) was necessary to suppress false positive voxels. For the final nuclei mask (Supplementary Fig. 8b, panel 6), it was joined with a simple intensity thresholded image ($I > 5{,}000$) using the union of the masks: that is, $proc_{DAPI} > 5{,}000$ OR ($proc_{DAPI} > thr_{sauvola}$ AND $proc_{DAPI} > 2{,}000$).

To achieve instance segmentation of the dataset (that is, into individual cells), each cell needs to be separated from its neighbors. Common procedures for instance identification are connected components or the detection of the cell center by a distance transform on the nuclei mask. This relies on an initial separation between the nuclei in the mask. However, this approach does not yield satisfactory results for densely packed tissue because most nuclei are abutting in the DAPI channel. Therefore, to robustly find the centers of the cells, we pooled the information from the the membrane mask and the nuclei mask to generate a mask with maximally separated nuclei. For this, the nuclei mask was eroded slicewise (that is, in each 2D plane) with a disk with a radius of 3 pixels. This eroded-nuclei mask is combined with the membrane mask by the difference operation: any pixels in the eroded-nuclei mask that are also in the membrane mask are removed (Supplementary Fig. 8b, panel 7).

Euclidian distance transformation (SciPy v.1.3.2) on the inverse of this combined mask yields a distance image coding the distance to the nearest nucleus center for each voxel (Supplementary Fig. 8b, panel 8). Hence, nuclei centers are expected to present as peaks in this image. Peak detection was done on the distance image to find the cell centers to be used as seeds for segmentation. Local maxima were identified using a maximum filter with an oblate ellipsoidal footprint ($diameters_{zyx} = [11, 19, 19]$ voxels). Because the voxel values in the distance image are discretized by the voxel grid spacing, it may happen that multiple peaks with the same height are contained in the search-region/footprint. To avoid this, a very small modulation was applied to the distance image by using a difference-of-Gaussians ($\sigma_1 = 2\,\mu m$, $\sigma_2 = 4\,\mu m$) filtered DAPI channel that is normalized between 1.00 and 1.01. This makes it likely the peak closest to the center of the cell will have the larger height in the modulated distance image. The peaks were then extracted using a threshold of 1.16 µm. In sum, seeds were detected in the center of DAPI-positive nuclei that are $\gtrsim$7.5 µm apart and >1.16 µm from the nearest membrane (Supplementary Fig. 8b, panel 9).

The nuclear space in the combined mask was then flooded from the peaks using a watershed operation on the distance image with the detected cell-center peaks as seeds (Supplementary Fig. 8b, panel 10) and masking voxels with a distance to the membrane smaller than 1.16 µm from the operation. To finish the cell segmentation, a second watershed operation was performed on the unmasked, smoothed ($\sigma = 1\,\mu m$) average membrane image using the outcome of the first watershed as seeds (Supplementary Fig. 8b, panel 11). To constrain the segments over the *z* axis, where membrane signal may be too weak to properly function as boundary, we adapted scikit-image's compact watershed to work with anisotropic data and used it for this final operation (compactness, 205). As a postprocessing step, only segments that are fully contained within the dataset mask are retained, where the dataset mask is obtained by thresholding the Gaussian-smoothed (in-plane, $\sigma = 16.6\,\mu m$) arithmetic average over all channels ($I > 1{,}000$).

Because our dataset incorporates channels with membrane markers and nuclear markers, we can even go beyond cellular resolution and extract specific compartments of the cell. This further increases the specificity of the marker intensity profiles of the cells. To achieve this, we perform a straightforward subsegmentation of the cells by marking the voxels on the segment boundaries and dilating this mask in-plane by 1 voxel (Supplementary Fig. 8b, panel 12). We define the nuclear compartment as the voxels of the segments that overlap with voxels in the nuclei mask (Supplementary Fig. 8b, panel 13). Voxels that are not in either the membrane or nuclear compartment are assigned to the cytoplasm compartment. Note that in this definition voxels can belong to both the membrane and nuclear compartment.

Deep learning segmentation. Integration with two deep learning models is provided for easy use within the STAPL-3D framework. For membrane enhancement, a 3D-UNET is used taken from the PyTorch[10] implementation. For nucleus segmentation, the StarDist[11] package is used.

Training datasets. To generate suitable mLSR-3D-specific training data for the deep learning models, we used a strategy of coacquisition of mLSR-3D data at the typical resolution and at a very high resolution at the same location (Supplementary Fig. 11a). Segmentations performed at the high resolution are then downsampled to create the training labels for the typical resolution mLSR-3D data. We acquired seven *z*-stacks with a ×63 objective (a roughly $160 \times 1{,}024 \times 1{,}024$ matrix with a $0.39 \times 0.13 \times 0.13\,\mu m^3$ voxel size) and ×25 objective (a roughly $70 \times 1{,}024 \times 1{,}024$ matrix size with $1.2 \times 0.33 \times 0.33\,\mu m^3$ voxels) at the same location. The locations of the *z*-stacks were chosen throughout the sample to have all populations represented in the training data. Segmentation of the high-resolution data was performed by a simple STAPL-3D pipeline. The weighted membrane mean was calculated and enhanced through membrane probability prediction with a generic 3D-UNET model trained on plant cells [confocal_unet_bce_dice_ds2x][10]. The membrane probability was thresholded ($P > 0.1$) to create a membrane mask. For creating a nucleus mask and the combined seed mask, we used the STAPL-3D method as described in Seed-detection segmentation. Individual seeds were then identified by connected component labeling of the seed mask, where seeds smaller than $50\,\mu m^3$ were removed and seeds larger than $5{,}000\,\mu m^3$ were subdivided further by eroding the mask of these labels (footprint, [5, 15, 15] voxels), again followed by connected component labeling. These subdivided labels were reinserted in the nucleus segmentation volume, replacing the original large labels, yielding the final nucleus segmentation. Watershed was then performed from the nuclei to the boundaries in the 3D-UNET predicted probability map to yield the cell segmentation.

The typical resolution image was then registered to the high-resolution image with elastiX[31] using a 12 d.f. affine model. The transformation matrix was

used to transform the channels from the typical resolution into the space of the high-resolution image, sampling the volume at the target mLSR-3D voxel spacing ($1.2\times0.33\times0.33\,\mu m$) with trilinear interpolation, yielding the training data. Finally, the high-resolution nucleus and cell segmentation were downsampled to the same resolution using nearest-neighbor interpolation, yielding the training labels ($n=18{,}627$ segmented cells).

Model training and prediction. For training the StarDist and 3D-UNET models for use with mLSR-3D data, the seven training volumes (resampled to a roughly $45\times400\times400$ matrix with resolution $1.2\times0.33\times0.33\,\mu m^3$) were split in five training and two validation sets.

The StarDist model was trained on the raw DAPI channel with the nucleus segmentation as labels. We followed the StarDist default settings (notably data normalization between the 1 and 99.8 percentiles; $n_{rays}=96$; grid, [1, 2, 2]; epochs, 400) using a patch size of [40, 96, 96] voxels. For data augmentation, random flips and 90° rotations were used, as well as intensity variation according to $I_{aug}=aI+b$, with a and b randomly chosen from intervals $a=[0.8, 1.2]$ and $b=[-0.1, 0.1]$. Training on a single NVIDIA Quadro RTX6000 graphics card took 41 h.

For prediction, we made use of the StarDist blocking system for big datasets, with adaptations to make it efficient for high-performance computing. Blocks ($n=462$; with $block_{size}=[Z_{max}, 1{,}024, 1{,}024]$; $min_{overlap}=[32, 128, 128]$; context, [0, 64, 64]) are predicted in parallel on the CPU, writing the predictions to file. Subsequently, predictions were concatenated to yield the merged nucleus prediction for the full dataset (Supplementary Fig. 11c, second row).

The 3D-UNET model for membrane probability estimation was trained on the weighted membrane mean with the cell segmentation as labels. We took the default parameters of the PlantSeg confocal boundary model (https://github.com/wolny/pytorch-3dunet/blob/62e10674c5cf9f44e252297203213b8c7f23c5f7/resources/3DUnet_confocal_boundary/train_config.yml) and adapted the training patches to our configuration (patch, [40, 80, 80]). Training on a set of four NVIDIA Quadro RTX6000 graphics cards took 9 h. Prediction (Supplementary Fig. 11c, fourth row) was performed on datablocks (as described in Parallelization) with patch, [64, 128, 128]; stride, [32, 100, 100] and $mirror_{padding}=[16, 32, 32]$.

With good nucleus and membrane predictions, we used a simple watershed segmentation model for STAPL-$3D^{DL}$ (Supplementary Fig. 11b). The StarDist prediction was taken as the final nucleus segmentation. For cell segmentation, we performed a straightforward watershed (no compactness) from the StarDist nuclei to the boundaries in the 3D-UNET prediction (Supplementary Fig. 11c, bottom row). The membrane and cytoplasm compartments were then generated as described in Seed-detection segmentation.

Airyscan 3D segmentation. Segmentation of the Airyscan data used the following STAPL-3D workflow. A cell mask was generated to separate the organoids from the background. For the 63× Airyscan data, the cell mask was found by thresholding ($I>200$) the Gaussian-smoothed ($\sigma=10\,\mu m$), weighted-average of the three channels ($w_{DAPI}=1.0$, $w_{FACTIN}=5.0$, $w_{MT}=5.0$). The nucleus mask was generated from the DAPI channel according to the procedure described in Seed-detection segmentation (grayscale opening; median in-plane filter of 5 px; sauvola window, [19, 75, 75]; sauvola $k=0.2$; exclusion threshold for 25×, $I<300$; exclusion threshold for 63×, $I<100$; inclusion threshold for 25×, $I>800$ and inclusion threshold for 63×, $I>200$).

For the 25× Airyscan data, membrane enhancement of the F-ACTIN channel was performed by prediction of boundaries with the HFK-trained 3D-UNET model. The UNET boundary probability was converted to a mask by thresholding ($P>0.5$), binary dilation and removing components smaller than $40\,\mu m^3$. For the 63× Airyscan data, ACME membrane enhancement (median filter parameter of 0.5 and neighborhood radius of 1.1) was performed, after which the planarity was thresholded ($I>0.001$) to yield the membrane mask.

The seed mask was then generated by combination of the nucleus and membrane mask and labeled for connected components. Seeds smaller than $20\,\mu m^3$ were excluded and those larger than $1{,}000\,\mu m^3$ were broken up by eroding their mask (footprint, [3, 9, 9]) and relabeled. Objects in the relabeled seed mask smaller than $10\,\mu m^3$ were removed.

A two-step watershed procedure was then used for cell segmentation, first from the seeds into the nucleus mask using the nucleus mask's distance transform and subsequently from the nuclei into the cell mask using the 3D-UNET boundary probability map or negative of the smoothed mean over channels for the 25× and the 63×, respectively.

To extract mitochondria, the MitoTracker channel was corrected for inhomogeneity (50 iterations; four fitting levels; seven b-spline controls points for 25× and nine b-spline controls points for 63×). A difference-of-Gaussians filter ($\sigma_1=0.1$ mm; $\sigma_2=0.2$ mm) was applied to the corrected image and then thresholded ($I>0.00005$) to yield the mitochondrial mask.

Automatic segmentation parameter tuning. With the labeled datasets that were generated for training the deep learning models, we automatically fine-tuned segmentation parameters using a Bayesian optimization process to find their optimal values (Supplementary Table 3). For this, we used the hyperopt Python package[26] and its default tree-structured Parzen estimator algorithm. For each iteration of the optimization, we segmented five separate regions (containing a total of 8,053 whole cells) using a different parameter set. Optimization was performed on eight parameters of the STAPL-3D segmentation pipeline for which the search space was set to:

- ACME neighborhood radius: 0.5 to 1.8 µm with increments of 0.01 µm
- nucleus mask erosion disk radius: 0 to 10 px with increments of 1 px
- peak detection footprint: (z), 1 to 21 px with increments of 2 px
- peak detection footprint: (yx), 1 to 101 px with increments of 2 px
- peak detection threshold: 0 to 2 µm with increments of 0.0001 µm
- membrane channel smoothing σ, 0 to 3.3 µm with increments of 0.33 µm
- membrane mask threshold; 0 to 0.01 with increments of 0.00001
- compactness; 0 to 500 with increments of 1.

To optimize the parameters, a single segmentation score is needed as a cost function. While the average Dice score (ADS) and F_β-score (which combines the precision and recall) each represent a view of the segmentation accuracy—at voxel level and segment level, respectively—they do not fully capture the segmentation performance. To assess segmentation in a single overlapping score, we combine both scores in an overall segmentation score (OSS):

$$\mathrm{OSS} = w_1\mathrm{ADS} + w_2F_\beta$$

where w_1 and w_2 are weight factors for each score. The choice for these weights may depend on the application and can be selected to give more emphasis on various types of segmentation error. In our experiments, we chose $\beta=1.5$, $w_1=0.5$ and $w_2=0.5$.

The negative of the OSS was used as the cost function for optimization. Iterations with a recall smaller than 0.4 were set to an OSS of zero to speed up the process. Optimization was performed with eight parallel processes and was stopped after 500 iterations. The set of parameters that resulted in the highest OSS was selected as optimal parameter set (Supplementary Table 3).

Imaris segmentation. We used the membrane-based segmentation method of the 'Cells' module in Imaris v.9.5.0. Including the nucleus estimation option deteriorated the segmentation; therefore it was not used. Segmentation parameters were adjusted based on visual inspection of segmentation of the labeled training datasets. The optimal workflow proved to be the local contrast method with a membrane size of 0.8, smallest cell diameter of 6 µm, minimum intensity threshold of 1,000 and a minimum quality threshold of 0.02. After segmentation, segments with a size smaller than 300 voxels were removed.

Ground truth dataset generation. To obtain an extensive ground truth validation dataset, we followed the approach described under Deep learning segmentation with ten separate coacquired volumes at two resolutions, high and typical, taken from diverse regions of the HFK. However, to segment the ground truth labels of the validation set, we designed a more complex pipeline to further maximize segmentation quality and minimize data curation efforts. In this pipeline, we determined membrane weights (used to generate the mean membrane channel) for each stack individually. Second, we took an iterative approach of the erosion-relabeling-expansion procedure for splitting seeds that were too large to represent single nuclei.

In one such iteration, the mask of large seed labels ($>341\,\mu m^3$) is eroded with a ball-shaped element of diameter D_{ero} voxels, attempting to break connections between neighboring nuclei. Then, the objects in the eroded mask are relabeled, after which very small labels are removed ($<0.341\,\mu m^3$). Finally, the individual labels are expanded with a ball-shaped element of slightly smaller radius $D_{exp}=D_{ero}-1$. Seeds that fall below the volume threshold of $341\,\mu m^3$ are then added to seed label volume.

We first performed two iterations with $D_{ero}=5$ and $D_{ero}=7$. We then combined the mask of remaining large labels with a low-threshold membrane mask, attempting to further break connections between nuclei, by removing additional voxels that have moderate membrane probability. We performed the erosion-relabeling-expansion procedure on this mask for seven iterations with $D_{ero}=[5, 7, ..., 17]$. Cell segmentation was then performed by watershed from the seeds to the 3D-UNET boundary estimate. Nuclear segments were obtained by clearing the space outside the nuclei mask.

The segmentation was thoroughly curated manually in ITK-SNAP (v.3.8.0) to resolve any remaining errors by (1) splitting, (2) merging or (3) removal for the most complex cases.

Accuracy metrics. The quality of STAPL-3D segmentation was assessed by comparing results to the ground truth on both the segment and voxel level. Assessment on the segment level provides information on the number of correctly identified cells and number of under- and oversegmented cells, which is mainly relevant for cell count statistics. Assessment on a voxel level provides information on the overlap and shape similarity, which is mainly relevant for the correct extraction of intensity, textural and morphological features for each individual segment.

To calculate a score for segmentation on a segment level, we assign every segment to either a truth label or the background based on largest overlap. A truth label with no assigned segments is seen as a false negative (FN). For a truth label

with one or more assigned segments, a single segment that has the largest overlap with the truth label (the main contributing segment) is counted as a true positive (TP), while any additional assigned segments are counted as false positives (FP). Using these numbers, we can calculate the precision and recall:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The precision and recall are representative of the oversegmentation and undersegmentation, respectively. While oversegmentation leads to a decrease in precision, undersegmentation leads to lower recall. To provide a single score for segmentation on the segment level, we use the F_β-score:

$$F_\beta = \frac{(1+\beta^2)\text{TP}}{(1+\beta^2)\,\text{TP} + \beta^2\text{FN} + \text{FP}}$$

For calculation of the F_β-score, $\beta = 1.5$ was used to consider recall more important than precision. Recall was considered more important in this case because we assume that undersegmentation has a higher negative effect on our subsequent analysis. Merge errors (undersegmentation) will mix intensities values found in different cells, while split errors (oversegmentation) will only extract the intensities from a subset of voxels of the cell. If we assume that the intensity distribution is more uniform within a single cell than over cell boundaries, the merge error can be considered worse than the split error.

For the assessment of segmentation performance on a voxel level, the Dice score is commonly used[32]:

$$\text{Dice} = \frac{2|T \cap A|}{|T| + |\text{A}|}$$

where T is a single truth label and A is the main contributing segment in the automatic segmentation as defined earlier. The Dice score is only calculated for truth labels that had a main contributing segment, as Dice scores of truth labels that have not been recalled (false negatives) are not representative for the performance of segmentation on a voxel level. All calculated Dice scores were then averaged to yield the ADS:

$$\text{ADS} = \frac{\sum_{k=1}^{n} \frac{2|T_k \cap \text{A}_k|}{|T_k| + |\text{A}_k|}}{n}$$

where n is the total number of labels in the ground truth with a main contributing segment, T_k is a single truth label with label k and A_k is the main contributing segment for T_k.

Segmentation accuracy comparison. To evaluate the different segmentation approaches, we scored them against the manually curated ground truth dataset (Supplementary Fig. 12a, left column). We included Imaris segmentation in this comparison, as a widely used external reference and benchmark for validation. Partial cells touching the border of the ground truth datasets were excluded before scoring, yielding a total of 14,717 cells distributed over ten datasets from diverse regions of the kidney. The datasets were segmented according to five different pipelines (Supplementary Fig. 12a, columns 2–6), specifically:

1. STAPL-3D: segmentation with classical image processing steps according to Seed-detection segmentation
2. STAPL-3D^{FT}: segmentation with the same STAPL-3D pipeline, but using the fine-tuned set of parameters as described in Automatic segmentation parameter tuning
3. STAPL-3D^{DL} (generic): segmentation using deep learning predictions of nuclei and membranes by models provided by their developers
4. STAPL-3D^{DL} (HFK-trained): segmentation using deep learning with models retrained on mLSR-3D data from the HFK as in Deep learning segmentation
5. Imaris: segmentation with the 'Cells' module as described in Imaris segmentation

Comparison of the $F_{1.5}$ scores for different segmentation methods was achieved by fitting the data to a linear regression model with random effect to account for variation between the ten different datasets, followed by a Tukey post hoc test. We compared morphological features, including cell volume, extent, equivalent diameters, minor and major axis length obtained from the different STAPL-3D segmentation modules to the ground truth for 7,063 true positive cells that were shared among all segmentation methods. Feature values of matching cell pairs were then used to calculate the mean positive and negative deviations from the ground truth values. Finally, an average of the mean deviations for the different cell shape features was calculated to represent an integrated measure of shape feature deviation (Supplementary Fig. 12c and Supplementary Table 3).

Zipping. Having parallelized the segmentation process for increased analysis speed and reduced memory footprint, the need arises to reassemble the blocks into a final combined segmentation volume without seams at the block boundaries. These seams are a consequence of trivial parallelization in processing the individual blocks (that is, without communication between the processes). They manifest through partial cells lying on the block boundaries that have been assigned different labels in different blocks. These doubly segmented cells may not perfectly match up over the boundary. These block-boundary segments need to be resegmented to complete the accurate segmentation of the full dataset. We refer to this correct reassembly of the datablocks as 'zipping' (Supplementary Fig. 8c). In short, it consists of identifying the segments lying on the boundaries, removing them and resegmenting that space. We aimed to design the procedure such that it requires minimal computational resources and expertise (fast, with a low memory footprint and without the need for communication between processes).

First, every segment label is made unique by sequentially relabeling the segments over all datablocks. Then, datablock pairs and quads (that is, the intersections of four datablocks) that exhibit overlap are queued for zipping.

Zipping sequence. In zipping, operations are not independent, but the procedure can be partially parallelized. To generate a fast implementation while ensuring computational processes cannot assign labels to the same datablock concurrently, we use the following sequence in computing the zipping (Supplementary Fig. 8c, top panel):

1. the zip for pairs on even zip-lines running over x (datablock pairs connected over the red solid lines in Supplementary Fig. 8c, top panel); each zip-line in a separate parallel process, each zip-pair in the zip-line processed sequentially
2. … odd zip-lines over x (red dotted lines)
3. … even zip-lines over y (green solid lines)
4. … odd zip-lines over y (green dotted lines)
5. the zip for quads on even/even zip-line intersects; (datablock-quads connected by the blue squares in Supplementary Fig. 8c, top panel); every quad in separate parallel process
6. … quads on even/odd zip-line intersects (cyan squares)
7. … quads on odd/even zip-line intersects (magenta squares)
8. … quads on odd/odd zip-line intersects (yellow squares)

Zipping procedure. The pairwise/quadwise zipping calculation is performed in subblocks of data (zip-block) around the zip-line/zip-quad, selected to completely contain the cells that lie on the seam (Supplementary Fig. 8c, bottom panel). A mask is generated of segments that touch the seam—extending the zip-block until the mask is contained within—in which peak selection and watershed segmentation is performed. In particular, for each zipping calculation on a zip-pair/zip-quad:

- the overlapping regions, that is the datablock margins ($N_{\text{margin}} = 64$ px), of the datablock-pair/datablock-quad of $2 \times N_{\text{margin}} = 128$ pixels (42.5 µm) are read into memory, while observing an additional margin of m multiples of the datablock margin ($m \times N_{\text{margin}}$ pixels): the zip-block. Initially, we set $m = 1$. For example, an initial zip-block (on a zip-line running over y) measures $106 \times 1{,}280 \times 256$ ($N_z \times N_{\text{block}} \times 2 \times 2 \times N_{\text{margin}}$), in which one half of the zip-block ($106 \times 1{,}280 \times 128$) contains segments from the one datablock, the other half from the other datablock. The seam is present in between
- the zip-mask is determined by selecting the voxels that belong to segments that touch the seam
- a check is performed to verify that this mask is fully contained within the selected margin; if any segments are larger (that is, touch the two relevant borders of the zip-block), m is incremented by 1 (that is, the margin is increased by 64 pixels)
- the zip-block resegmentation is done identically to the datablock segmentation, but the procedure is constrained to the zip-mask: seeds/peaks are selected from the zip-mask only and the watershed-fill is constrained to the zip-mask
- the zip-block segments are inserted in the datablock-pair/datablock-quad segmentations, ensuring unique labeling of each cell

Finally, after computing all zips, the segmentation datablocks are assembled, without their margins, in a single HDF5 datafile (and Imaris v.5.5 file format for visualization).

Feature extraction. The segmentation that we perform is aimed at extracting information from specific cellular compartments from large-scale (mLSR-3D) imaging datasets to perform spatio-phenotypic patterning of tissues. Therefore, over 800 features were extracted for each cell (Supplementary Fig. 8d and Supplementary Table 4). Pearson correlation coefficients were computed between all features over all the cells in the dataset and visualized with the R package corrplot[33] to screen the extensive feature set for discriminative features. We selected the most relevant roughly 20 features for our subsequent clustering and pseudotime analysis. The information that is derived from the segmentation is coded in a feature vector (see 'Feature set' below) in which four types of feature are distinguished: intensity, textural, morphological and spatial features

(Supplementary Fig. 8d, left panel). Additionally, we defined new morphological and spatial features specific to biology and to our application to HFK and its associated Wilms tumor. This feature extraction analysis stage yields a $N \times M$ (cells × features) matrix that is exported for use in downstream analysis.

Feature set. Features were computed for the full cell as well as for the membranal and nuclear subsegments (Supplementary Fig. 8d, left panel). As intensity features, the mean, median, minimum and maximum intensity over the (sub)segment's constituent voxels were computed for each channel. For the texture features, the range and variance were computed over the segment intensities, as well as the weighted center of mass and various forms of the weighted image moments. The morphological feature set consisted of volume, center of mass, extent, Euler number, image moments (raw, central, normalized) and inertia tensor eigenvalues (major/minor axis length). These basic cellular and compartment-specific intensity and morphology metrics are extracted using scikit-image's regionprops module (https://scikitimage.org/docs/dev/api/skimage.measure.html#skimage.measure.regionprops), adapted to return the minor/major axis length for 3D segments and the region's median value and variance.

Further, we calculate a measure of ellipsoidal shape from the inertia tensor eigenvalues by calculating the fractional anisotropy (FA)[34] that ranges from 0 to 1 for spherical ellipsoids to very elongated ellipsoids (formally, with only one nonzero eigenvalue). Furthermore, a measure of cell polarity was calculated as the distance between the segment's center of mass and the DAPI-channel intensity-weighted center of mass, divided by the segment's major axis length (Supplementary Fig. 8d, top-right). Finally, the spatial position with respect to the coordinate system of the sample (that is, peripheral-to-central) was captured by a distance-to-edge feature (Supplementary Fig. 8d, bottom-right). Distance-to-edge was defined as the Euclidian shortest in-plane distance from the cell's center of mass to the boundary of the sample mask.

Feature postprocessing. The feature extraction is performed through distributed processing on the individual datablocks. The cells × features matrices of each block are concatenated over cells. At this stage, a selection of features and cells is made. Specifically, the final feature set (19 features) to enter in the developmental profiling analysis is selected, taking the values from the appropriate subsegments:

- median marker intensities for DAPI, PAX8 and SIX2 (from nuclear subsegments)
- median marker intensities for NCAM1, CADH1, CADH6 and F-actin (from membrane subsegments)
- KI67 classification for each cell (see 'KI67 quantification')
- fractional anisotropy, cell volume, extent, major axis length, minor axis length and equivalent diameter (from full segments)
- distance-to-edge and center of mass *z*, *y* and *x* (from full segments)
- polarity (compound from full segments and nuclear subsegments)

In addition,

- for any cells with a very small nucleus subsegmentation (<6.6 µm³; 50 voxels), the values of the nucleus-derived intensity features are replaced with the values of the full segment intensity features
- any cells that touch the border of the dataset are discarded (that is, these cells may only lie in the dataset partially)
- cells for which any of the selected numerical features (that is, 18 features, all except KI67) are zero are discarded
- cell duplicates represented in multiple datablocks are discarded

Parallelization. Our parallel computing implementation of the segmentation procedure enabled fast and efficient analysis of the very large datasets generated for this study. For this work, the datasets were subdivided over *xy* with a blocksize of $N_{block} = 1{,}280$ pixels into ~200 blocks ($blocksize_{zyx} = Nz \times 1{,}280 \times 1{,}280$ voxels) with an overlap/margin of $N_{margin} = 64$ pixels ($margin_{zyx} = 0 \times 64 \times 64$ voxels) that were segmented in parallel. This choice of block size is adpated to our particular combination of data matrix and compute devices, but arbitrary block sizes can be chosen to suit other configurations. A single block (of this particular size) was segmented in ~1 h with a memory requirement of 40 GB (for the membrane enhancement step). On our 1,844-node high-performance computing cluster, the seed detection segmentation could therefore—in times of good compute node availability—be performed in ~1 h, while at prediction time the STAPL-3D deep learning variant is an order of magnitude faster still, either if performed massively parallel on the CPU or using ~8 cards of the HPC's GPU partition. The block-zipping was done in ~1.5 h (3 GB) and the feature extraction in ~1 h (10 GB). On our high-end workstation (HP Z8 G4 with dual Intel Xeon Gold 5122 3.6 GHz processors, 1 TB RAM, 3×2 TB SSD + 6×5 TB HDD in RAID-5, and an Nvidia Quadro P4000 8GB graphics card), the STAPL-3D segmentation was performed in ~9 h; block-zipping took ~3 h; and feature extraction of the large feature-set added another ~9 h.

Quality assurance. For each step in the STAPL-3D pipeline, a summary pdf report of the input and output data is provided, as well as detailed output of intermediate steps on request. Data correction reports include images of before and after correction and estimated inhomogeneity (Supplementary Fig. 9a,b). Segmentation reports give overview statistics, as well as visual impressions of the quality of intermediate masks and final segmentations (Supplementary Fig. 9c,d). The reports can be used as a check for data quality and as guidance for subsequent analysis steps. Furthermore, with this output it is straightforward to flag any potential issues in the analysis at an early stage, which can be challenging for large datasets where instant inspection is difficult, particularly when analyzed on a compute cluster.

Cluster membership and pseudotime. To gain insight into the developmental processes in HFK and their potential parallels in Wilms tumor, we devised a machine-learning strategy of automated assignment of cells to spatio-phenotypic populations and pseudotime order. This strategy applies established tools from the scRNA-seq field[35] to our adequately rich imaging data to achieve clustering and pseudotime estimation on the HFK data. Because the number of cells in our dataset is prohibitively large for direct calculation of these measures, we train classifiers on a subset of the cells and predict cluster membership and pseudotime on the full dataset. We then also use these same classifiers to assign cell types and pseudotime ordering to the cells in the Wilms tumor.

For the HFK sample, the data in the cells × features matrix were log transformed and standardized by scaling each column to zero-mean and unit-variance. Fifty-thousand cells were randomly selected from the data matrix to serve as training set. We further select 12 discriminative features (intensities—PAX8, SIX2, NCAM1, CADH1, CADH6, F-actin; morphological—extent, fractional anisotropy, major axis length, minor axis length; spatial—distance-to-edge, polarity) for projecting the data into three dimensions with the UMAP algorithm[21] (using the defaults from umap-learn (v.0.4.0rc1), but with three components and 30 nearest-neighbors).

In scanpy (v.1.4.5.2.dev33+g8d26ad5e)[35], a neighborhood graph ($N_{neighbors} = 30$) was built from the UMAP after which clustering was performed with the Leiden algorithm[36] using a resolution parameter of 1.50.

With the cells in the training set assigned to a population, we use scikit-learn (v.0.22.2.post1)[37] to train a support vector classifier (SVC) for clustering using the UMAP embedding of the feature vectors as input data and the clusters as data labels. The UMAP transform as derived from the training set was applied to the feature vectors of all cells in the data matrix, after which the classifiers were used to predict population membership for every cell in the dataset.

For the Wilms tumor sample, the intensities of the eight channels were rescaled to the range of the HFK sample by taking the ratio of the intensity of characteristic structures of interest for each channel. Similar to the HFK, log transformation and standardization was applied to each feature in the data matrix. Each cell in the tumor dataset was then assigned to one of the populations as found in the HFK through classification with the SVC.

To leverage the benefit of the detailed spatial information in our data, we use the cell IDs to backproject (see 'Backprojection and image visualization'), among others, the cluster assignment of each cell into the original imaging space. We can then color-code and select the cells in image space according to population. Therefore, we can accurately verify sensible cluster assignment by expert evaluation.

Furthermore, we measured the performance of the SVC and assessed misclassification behavior by means of a confusion matrix (Supplementary Fig. 15c). The reference dataset consisting of a subset of 50,000 cells with assigned cell types was split into training (90%) and test (10%) datasets. The training dataset was used to train a SVC for assigning cell types using the UMAP transform. Then, the SVC was used to predict the cell types on the unseen test dataset. Predicted values were plotted against true labels of the test dataset in a relative confusion matrix to assess the percentage of error for each class.

On the basis of the backprojection of the clustering, the clusters were reassigned to one of 11 known populations comprising nephrogenesis, collecting system and interstitium. Using only the cells in the nephrogenic clusters, a new UMAP (two components, 30 nearest-neighbors) was generated and a pseudotime trajectory was calculated.

For pseudotime ordering with scanpy, a diffusion map was calculated with 15 diffusion components on the training set and diffusion pseudotime[38] was estimated (without branching) from the first five components starting from a SIX2-high cell (that is, picked randomly from all cells in the 99th percentile of the SIX2 intensity feature).

Similar to cluster assignment, the pseudotime trajectory was generalized to all cells in the HFK dataset, as well as the Wilms tumor, through a support vector regressor in scikit-learn.

Three developmental paths were defined for the nephrogenic process, ending in the distal tubule (CM—RV/SSB—distal tubule), the proximal tubule (CM—RV/SSB—proximal tubule progenitor—proximal convoluted tubule) and the glomerulus (CM—RV/SSB—glomerulus). Markers' evolution over pseudotime was analyzed by sorting the nephrogenic part of the data matrix according to first cluster membership, then pseudotime, and plotting the trajectories using a moving average over 20,000 cells.

For joint analysis of the HFK and Wilms tumor data, the cells × features matrices were concatenated. The training set for generating a joint UMAP and population and pseudotime classifier was composed of 50,000 cells from

the HFK and 50,000 cells from the Wilms tumor dataset. For this analysis, the distance-to-edge feature was omitted from the features that were used to generate the joint UMAP, because it is not informative in the tumor. Leiden clustering was performed on the joint UMAP with a resolution parameter of 0.4 to yield six clusters. From these six clusters, cluster no. 1 was isolated and a more fine-grained subclustering was achieved by generating a two-component UMAP with Leiden clustering at resolution of 0.7.

Backprojection and image visualization. Retention of the identity of every segmented cell allows visualization of computed features or any cell-specific derived quantity (for example, cluster membership) in the spatial coordinate frame of the dataset (backprojection). This provides the important advantage of interpreting results while considering the spatial position and environment of the cells. Datasets were visualized in Imaris imaging software v.9.5.0 (Bitplane). Stitched data were converted with Imaris File Converter (v.9.5.0). For Imaris visualization of STAPL-3D preprocessed data, segmentation and backprojected metrics, data were converted to unsigned 16-bit integers and written to the Imaris v.5.5 HDF5 file format, as a single channel per file. Composite files were created for data, segmentation and backprojection overlays by creating a file with the Imaris file-structure with symbolic links to the separate datasets. For visualizing segmented and backprojected data, we overlayed these added channels with the DAPI or membrane sum channel in blend mode.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data are publicly available. Processed results (that is, cells × features matrices and clustering and pseudotime results) and imaging data are made available through public repositories for which the links are posted on the STAPL-3D GitHub page.

Code availability

We provide the STAPL-3D framework as a Python package on github (https://github.com/RiosGroup/STAPL3D).

References

27. Dekkers, J. F. et al. High-resolution 3D imaging of fixed and cleared organoids. *Nat. Protoc.* **14**, 1756–1771 (2019).
28. van Ineveld, R. L., Ariese, H. C. R., Wehrens, E. J., Dekkers, J. F. & Rios, A. C. Single-cell resolution three-dimensional imaging of intact organoids. *J. Vis. Exp.* **2020**, 1–8 (2020).
29. Yushkevich, P. A. et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage* **31**, 1116–1128 (2006).
30. Sauvola, J. & Pietikäinen, M. Adaptive document image binarization. *Pattern Recognit.* **33**, 225–236 (2000).
31. Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. W. Elastix: A toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* **29**, 196–205 (2010).
32. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).
33. Wei, T. & Simko, V. corrplot. R Package, v. 0.84 (2017).
34. Basser, P. J. & Pierpaoli, C. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J. Magn. Reson. Ser. B* **111**, 209–219 (1996).
35. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
36. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, (2019).
37. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
38. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

Acknowledgements

We are grateful for the technical support from the Princess Máxima Center for Pediatric Oncology and Zeiss for imaging support. We acknowledge the Gynaikon Clinic in Rotterdam for their efforts to provide the human fetal material and the laboratory of Hans Clevers and the Hubrecht Organoid Technology (HUB, www.hub4organoids.nl) for access to the breast cancer organoid biobank. We also acknowledge the Utrecht Bioinformatics Center High Performance Computing Facility for data processing infrastructure. We thank R.R. de Krijger for useful discussions. All the imaging was performed at the Princess Máxima Imaging Center. This work was financially supported by the Princess Máxima Center for Pediatric Oncology and St. Baldrick's Robert J. Arceci International Innovation award. J.F.D. is supported by a VENI grant from the Netherlands Organisation for Scientific Research (NWO). A.C.R is supported by an ERC-starting grant 2019 project no. 804412.

Author contributions

R.L.v.I. and M.K. contributed equally. M.A. and S.d.B. contributed equally. R.L.v.I. developed mLSR-3D and performed microscopy. M.K. developed STAPL-3D and performed the computational methods. R.L.v.I. and M.K. analyzed the data. S.d.B. assisted with computational analysis. C.M.M. assisted with microscopy. M.B.R. performed mLSR-3D imaging of breast and neuronal tumor tissue, rendered data and made videos. E.J.v.V. performed mLSR-3D imaging of breast and neuronal tumor tissue. M.A. performed quality control and computational analysis. J.F.D. provided organoid and xenograft material. H.R.J. assisted with sample preparation. F.L.B. provided microscopy support. R.H. provided the NCAM nanobody used in this study. S.M.C.d.S.L provided human fetal material. J.F.D., M.A., E.J.v.V., S.d.B., F.L.B. and J.D. provided critical feedback on the work. R.L.v.I., M.K. and A.C.R. designed the study and wrote the manuscript with support from E.J.W., and A.C.R. supervised this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00926-3.

Correspondence and requests for materials should be addressed to A.C.R.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Anne C. Rios

Last updated by author(s): 28/01/2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Zeiss Zen Black Edition (v2.3)
Data analysis	Zeiss Zen Blue Edition (v2.6), Arivis V4D (v3.1.3), Imaris (v9.3, v9.5 or v9.6), Imaris File Converter (v9.5.0), CellProfiler (v3.1.9), N4 bias correction algorithm (v1.2.4), ACME (64-bit Linux binary), ilastik (v1.3.2), ITK-SNAP (v3.8.0), scikit-image (v0.16.2), scikit-learn (v0.22.2.post1), SciPy (v1.3.2), hyperopt Python-package (v0.2.2), Corrplot R package (v0.84), Scanpy Python package (v1.4.5.2.dev33+g8d26ad5e), UMAP-learn python package (v0.4.0rc1), Leiden clustering algorithm (v0.7.0), UNET-3D: pytorch-3dunet (v1.2.12), StarDist: stardist (v0.6.0), Elastix: simpleitk (v1.2.0rc2.dev1162+g2a79d), tensorflow (v2.2.0), keras (v2.3.1), pytorch (v1.4.0), napari (v0.3.5), pyyaml (v5.3.1), pandas (v1.1.0), numpy (v1.19.1), h5py (v2.10.0), nibabel (v3.1.1), matplotlib (v3.3.0), czifile (v2019.7.2), pypdf2 (v1.26.0), plotly (v4.6.0), holoviews (v1.13.2), datashader (v0.10.0), bokeh (v2.0.1), reshape2 (1.4.4), xlsx (0.6.4.2), dplyr (1.0.2), ggplot2 (3.3.2), readr (1.4.0), plyr (1.8.6), emmeans(1.5.3), lme4 (1.1-25).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data is publicly available. Processed results (i.e. cells ⊡ features matrices and clustering and pseudotime results) and imaging data are made available through public repositories for which the links are posted on the STAPL-3D GitHub page.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by the amount of cells present in the imaged volumes.
Data exclusions	No data was excluded.
Replication	All microscopy and segmentation has been successfully repeated several times with similar results.
Randomization	Randomization was not applicable and therefore not performed.
Blinding	Blinding was not used during this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Antibody / dye Vendor Catalog #
1 anti-B-CAT-AF568 Abcam ab201823
2 DAPI ThermoFisher D3571
3 anti-KI67-AF488 eBioscience 53-5698-82
4 anti-KI67-eFluor660 eBioscience 50-5698-82
5 Phalloidin-AF555 Invitrogen A-34055
6 Phalloidin-AF647 ThermoFisher A-22287
7 Phallodin-ATTO 532 Sigma-Aldrich 49429
8 Phalloidin-ATTO 700 Sigma-Aldrich 79286
9 anti-NCAM1-HiLyte Fluor 488 QVQ Holding BV Q55c-488
10 anti-NCAM1-HiLyte Fluor 555 QVQ Holding BV Q55c-555
11 SiR-actin Tebu-bio SC001
12 SiR700-actin Tebu-bio SC013

13 TO-PRO-3 ThermoFisher T3605
14 anti-CDH1 ThermoFisher 13-1900
15 anti-CDH1 BD Bioscience 610182
16 anti-CDH3 BD Bioscience 610228
17 anti-CDH6 R&D Systems AF2715
18 anti-CD31 Abcam ab134168
19 anti-CD146 Abcam ab75769
20 anti-HER2 ThermoFisher MA514509
21 anti-Jagged1 R&D systems AF1277
22 anti-Iba1 NovusBio NB100-1028SS
23 anti-K5 Biolegend 905903
24 anti K8+K18 Abcam ab17139
25 anti-KI67 ThermoFisher 14-5698-80
26 anti-GPCR LGR6 Abcam ab126747
27 anti-MAP2 Biolegend 822501
28 anti-PAX8 Bioconnect GTX31119
29 anti-SIX2 Proteintech 11562-1-AP
30 anti-alpha SMA Abcam ab7817
31 anti-Vimentin Merck CBL202
32 anti-WT1 Abcam ab89901
33 anti-mouse AF488 ThermoFisher A-21131
34 anti-mouse AF488 ThermoFisher A-21202
35 anti-rat AF488 ThermoFisher A-21208
36 anti-sheep AF488 ThermoFisher A-11015
37 anti-rabbit AF488 ThermoFisher A-21206
38 anti-mouse AF514 ThermoFisher A-31555
39 anti-rabbit AF514 ThermoFisher A-31558
40 anti-mouse CF514 Biotium 20483
41 anti-rat ATTO532 Rockland 612-153-120
42 anti-mouse AF555 ThermoFisher A-21137
43 anti-mouse AF555 ThermoFisher A-31570
44 anti-rat AF555 ThermoFisher A-21434
45 anti-rat AF555 Abcam ab150154
46 anti-sheep AF555 ThermoFisher A-21436
47 anti-rabbit AF555 ThermoFisher A-31572
48 anti-mouse AF568 ThermoFisher A-10037
49 anti-rabbit AF568 ThermoFisher A-10042
50 anti-rat CF568 Biotium 20092
51 anti-rat AF568 ThermoFisher A-11077
52 anti-mouse AF594 ThermoFisher R37121
53 anti-mouse AF594 ThermoFisher A-21203
54 anti-rat AF594 ThermoFisher A-21209
55 anti-goat AF594 ThermoFisher A-11058
56 anti-rabbit AF Plus 594 ThermoFisher A-32754
57 anti-mouse AF647 ThermoFisher A-31571
58 anti-mouse AF647 ThermoFisher A-21239
59 anti-rat AF647 ThermoFisher A-21247
60 anti-rabbit AF647 ThermoFisher A-31573
61 anti-mouse AF660 ThermoFisher A-21055
62 anti-mouse AF633 ThermoFisher A-21052
63 anti-rat AF633 ThermoFisher A-21094
64 anti-sheep AF633 ThermoFisher A-21100
65 anti-chicken AF647 ThermoFisher A-21449
66 anti-mouse AF700 ThermoFisher A-21036

Validation

Every antibody was validated by single-antibody staining.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Mus Musculus, NOD.Cg-PrkdcSCID Il2rgtm1Wjl/SzJ, Female, 6-8 weeks of age

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Breast tumor tissue was obtained from xenograft mouse models approved by the Animal Welfare Committee of the Princess Máxima Center and established in compliance with both local and international regulations

Note that full information on the approval of the study protocol must also be provided in the manuscript.